Биомедицинская химия, 2010 том 56, вып. 1, с. 82-89.

БИОИНФОРМАТИКА

УДК 577.2, 577.1

БАЗЫ ЗНАНИЙ В ПОСТГЕНОМНОЙ МОЛЕКУЛЯРНОЙ БИОЛОГИИ

А.В. Лисица[1]*, Б.В. Шилов[2], П.А. Евдокимов[1], С.А. Гусев[1,3]

[1]Учреждение Российской академии медицинских наук Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича РАМН, 119121, Москва, ул. Погодинская, д. 10; тел.: +7(499)246-37-31; факс +7(499)245-08-57; эл. почта: fox@ibmh.msk.su
[2]ГОУ ВПО Сибирский государственный медицинский университет Росздрава.
[3]Научно-исследовательская компания "Криптом"

Базы знаний могут стать эффективным инструментом, существенно повышающим качество решения задач исследовательского поиска в молекулярной биологии, способствующим развитию новых методов обучения и прогнозированию развития биомедицинской сферы. Развитие основанных на знаниях технологий должно приблизить очередной "сдвиг парадигмы" в науках о жизни за счет интеграции научно-исследовательских коллективов при системном решении актуальных задач постгеномной эры. В дискуссионной статье рассматривается концепция создания базы знаний на основе алгоритмической обработки данных при обращениях исследователей к молекулярно-биологическим Интернет-ресурсам.

Ключевые слова: база знаний, формализация знаний, пертинентность, масс-спектрометрия, протеом человека.

ВВЕДЕНИЕ. Сложно представить современную молекулярную биологию без электронных хранилищ информации - баз данных. В ресурсах NCBI размещается информация о последовательностях ДНК и РНК, а также о целых геномах [1]. Последовательности аминокислотных остатков белков хранятся в базе данных UniProt [2], а сведения о пространственной структуре - в базе данных PDB [3]. Кроме ресурсов, содержащих данные о структуре биомакромолекул, было разработано большое количество систем, описывающих функцию белков, например, в системе онтологий GeneOntology содержится информация о внутриклеточной локализации и клеточных процессах [4], в KEGG - о метаболических путях [5]. В последние 10 лет, в связи с развитием протеомики созданы ресурсы для хранения масс-спектрометрических данных об идентификации белков в плазме крови и различных тканях [6].

Несмотря на широкое применение баз данных, сегодня практически не уделяется внимание информационно-аналитическим системам, называемым базами знаний. Причину этого определенно назвать затруднительно, но возможно, неприятие баз знаний является затянувшейся реакцией на так и не оправдавшиеся надежды, которые возлагались на них в 70-80-е годы прошлого века [7]. В то время полагали, что экспертные знания удастся формализовать вручную и внести в виде формальных правил в аналитическую систему. В реальности оказалось,

* - адресат для переписки

что эффективного способа формализации знаний на основе опроса эксперта не существует, а значительные трудозатраты приводили к созданию наивных систем, отражающих самые очевидные закономерности. Тривиальные закономерности не позволяли генерировать интересные гипотезы: как следствие, принципиальные недостатки первоначальных баз знаний не позволили развиться аналитическим инструментам следующего уровня сложности - экспертным системам.

Развитие информационных технологий, прежде всего развитие Интернета, кардинальным образом изменило ситуацию в отношении доступности знаний. Формализация знаний теперь может производиться не вручную, а с использованием технологий machine learning (термин предложен K. Samuel) в ходе автоматического интеллектуального анализа существующих информационных ресурсов. Предметом анализа могут стать типовые сценарии работы экспертов с молекулярно-биологическими ресурсами, а его результатом - формализованные знания экспертов в форме, доступной для статистической и иной обработки компьютерными программами. Обращение к таким знаниям позволяет существенным образом повысить эффективность работы обычного пользователя, не являющегося экспертом в специальной области.

1. ОТЛИЧИЕ БАЗ ЗНАНИЙ ОТ БАЗ ДАННЫХ.

С технической точки зрения существует единственное различие между базами данных и базами знаний. Информационные объекты в составе базы данных индивидуализированы и разобщены, тогда как в составе базы знаний объекты объединены - между ними устанавливаются взаимосвязи и закономерности, описывающие предметную область. При этом важно, что в базе знаний содержится универсальный алгоритмический аппарат, позволяющий сопоставлять объекты друг с другом и оценивать степень их ассоциативной связности.

Указанное техническое отличие приводит к возникновению качественно новых свойств у баз знаний по отношению к базам данных (табл.). Прежде всего, смещается "центр тяжести": если при создании базы данных во главу угла ставится фактографическая информация, то при функционировании базы знаний в центре внимания находится пользователь [8]. Задача базы знаний заключается в организации информационного массива таким образом, чтобы пользователю предоставлялись только факты, имеющие непосредственное отношение к текущей решаемой проблеме. Свойство аналитической системы выявлять информацию, отвечающую интересам пользователя, называется пертинентностью (см. табл.). Обычно, в результате запроса к базе данных пользователь получает лишь релевантную информацию (степень соответствия результатов информационного поиска заданному запросу), формально совпадающую с ключевыми словами запроса. В отличие от релевантности, пертинентность - субъективное понятие, обозначающее степень соответствия результатов информационного поиска ожиданиям (задачам) пользователя поисковой системы.

Таблица. Сопоставление двух типов информационных систем.

Характеристика	База данных	База знаний
Предмет анализа	Фактографические данные	Пользователь
Критерий соответствия запросу пользователя	Релевантность	Пертинентность
Режим взаимодействия с пользователем	Ожидание запроса	Адресный маркетинг

Различие между релевантностью и пертинентностью можно пояснить на примере известной поисковой системы Google. Успех проекта Google связан с такой особенностью: в первую очередь в список найденных результатов выводятся страницы, на которые в прошлом обращало внимание большее число пользователей [9]. Этот элемент знаний, включенный в алгоритм Интернет-поиска, позволил Google буквально в течение нескольких месяцев стать более популярным, чем известные поисковики - AltaVista и Yahoo, в которых использовался традиционный поиск по релевантности [10]. Сейчас принцип ранжирования Интернет-страниц по частоте цитирования используют все существующие системы Веб-поиска.

Возможность выявления пертинентной информации позволяет базе знаний активно взаимодействовать с пользователем (табл.). Это также является существенным отличием от баз данных, которые ориентированы на ожидание запроса пользователя. Например, при выборе определенной книги в электронном магазине AMAZON.COM система автоматически предлагает дополнительную информацию о том, какие ещё книги интересуют покупателей выбранного романа. Хотя эту информацию пользователь не запрашивал, бизнес-логика сайта автоматически осуществляет адресный маркетинг.

Основанные на знаниях технологии постепенно интегрируются в систему ресурсов Entrez. В 2009 году появились так называемые сенсоры, пример работы одного из которых показан на рисунке 1 [11]. Сенсор предназначен для анализа запроса пользователя. Если в запросе присутствует проблематика гриппа H1N1 ("свиной" грипп), то на странице результатов поиска автоматически выводится информация о наличии новых сиквенированных геномов этого вируса.

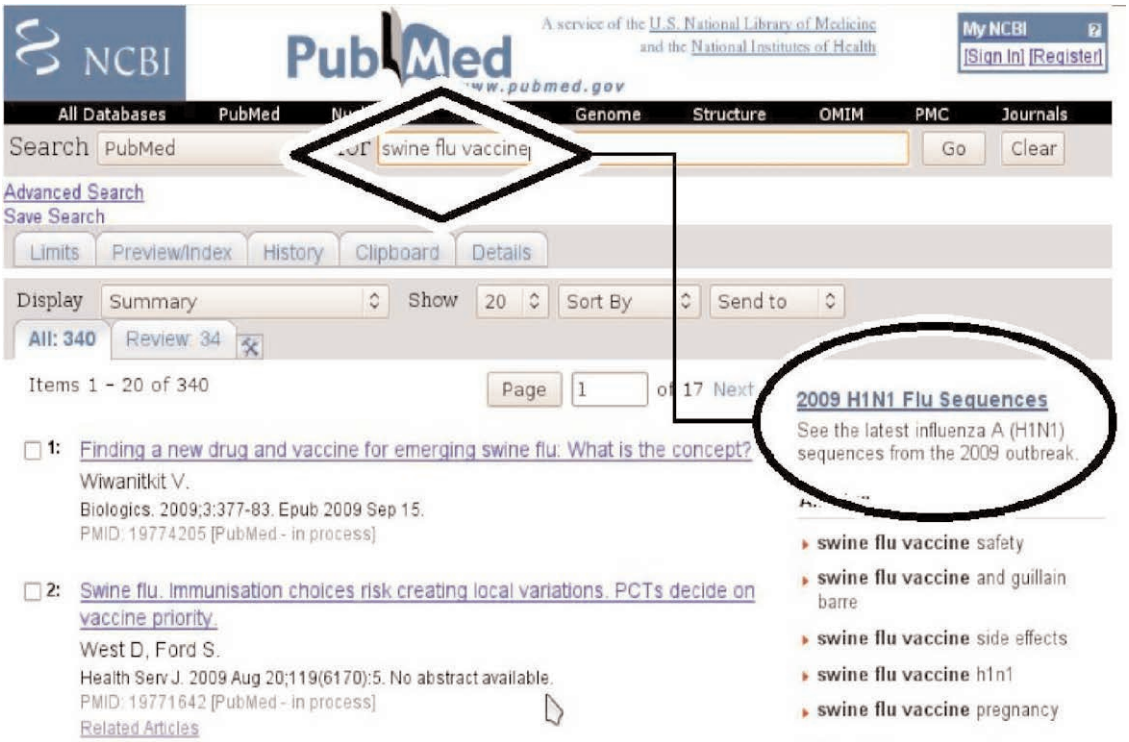

Рисунок 1.
Сенсор в системе PubMed предоставляет пользователю дополнительную информацию о расшифрованном геноме вируса H1N1, если предметом поискового запроса является “свиной” грипп.

2. КОНЦЕПЦИЯ БАЗЫ ЗНАНИЙ ДЛЯ РЕШЕНИЯ МОЛЕКУЛЯРНО-БИОЛОГИЧЕСКИХ ЗАДАЧ.

Основанные на накоплении знаний технологии могут быть реализованы с использованием существующей системы информационных ресурсов в области молекулярной биологии. Возможно, базы знаний целесообразно применять и в других областях науки и техники, однако ни одна из отраслей не может сравниться с биомедициной по объему и разнообразию электронных источников информации. Кроме того, существуют как высокоструктурированные информационные ресурсы, такие как базы данных белков и генов, так и ресурсы описательные, например, коллекция рефератов MEDLINE.

Современная работа с литературными источниками в большинстве случаев осуществляется посредством сети Интернет. Интернет используется для получения информации о структуре и функции как биомолекул, так и взаимодействующих с ними биологически активных химических соединений. Таким образом, отслеживая последовательность обращений к соответствующим ресурсам можно получить представление о том, над какой научной проблемой работает учёный.

Технически современный Веб-сервер осуществляет протоколирование всех действий пользователя, включая Интернет-адрес, время обращения к ресурсу, адрес Веб-страницы, с которой перешел пользователь, содержание поисковых запросов, выбранные гиперссылки и др. Обычно в отношении этой информации применяется политика конфиденциальности, которая, по сути, обозначает, что протоколы доступны для анализа только коллективу разработчиков Веб-ресурса [12] для решения узкого круга задач по оптимизации пользовательского интерфейса.

Использование результатов такого мониторинга может быть намного более эффективным в рамках концепции базы знаний. В частности, предоставленные пользователем данные могут использоваться для выявления пертинентной информации, отражающей специфику решаемой пользователем научно-исследовательской задачи. Для этого должна производиться обработка статистически значимого количества протоколов обращений к различным молекулярно-биологическим ресурсам. В результате этой обработки можно будет установить совпадающие у нескольких пользователей профили, своего рода "шаблоны" интеллектуальной активности. В дальнейшем, если один из пользователей столкнется с похожей проблемой, то база знаний автоматически выведет его на соответствующий шаблон, за счет чего расширится восприятие проблематики. Таким образом, неявные (неотчуждаемые от человека) знания большого количества пользователей могу быть формализованы, стать явными, доступными к распространению среди других исследователей.

Конечно, технологии накопления и обработки знаний могут использоваться намного более широко, чем изложено выше. При этом не меняется основной принцип базы знаний - пользователь получает информацию в переработанном с учётом предыдущего опыта виде. Изложенную концепцию можно проиллюстрировать на примере подхода к проблеме стандартизации результатов протеомных экспериментов.

3. ПЕРСПЕКТИВЫ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ БАЗ ЗНАНИЙ.

В состав Международной организации "Протеом человека" входит комитет по стандартизации, задачей которого является разработка форматов для хранения результатов протеомных экспериментов [13]. В настоящее время разработка форматов завершена, и на их основе Европейский институт биоинформатики создал центральный репозиторий для хранения масс-спектрометрических данных об идентификации белков PRIDE [6]. Предполагается, что исследовательские группы, выполняющие протеомные проекты, будут размещать свои результаты в системе PRIDE. Опыт функционирования системы PRIDE в течение 4-х лет показал, что в действительности информация поступает редко, в большинстве случаев, в связи с намерением исследователей опубликовать свои результаты

в журнале, поддерживающем стратегию стандартизации. Члены комитета по стандартизации исследовали причину низкой активности и выявили, что она заключается отсутствии мотивированности исследовательских групп в публикации масс-спектрометрических данных в системе PRIDE.

Решение этой проблемы возможно за счёт использования баз знаний, работающих по следующей схеме. Размещая перечень идентифицированных белков на общедоступном ресурсе, экспериментатор получает постоянную обратную связь со всеми другими исследователями, причем совершенно необязательно работающими в области протеомики. Не представляет сложности выявить такие профили знаний, где содержится наименование определенного белка либо в статье, либо в записи такой базы данных как UniProt, ProteinAtlas или NCBI protein. Обобщив указанные профили, можно получить усредненный тренд (поддающаяся численному представлению тенденция), отображающий, в каком контексте большинство современных исследователей "соприкасаются" с каждым белком из идентифицированного перечня. Разместивший свои данные в системе PRIDE исследователь получает, во-первых, материал для обсуждения в статьях своих результатов под различными углами зрения. Во-вторых, для него открываются широчайшие возможности для установления сотрудничества, в том числе, с целью написания совместных статей на стыке разных областей научной компетенции. В итоге, размещение данных на общественном ресурсе прямым образом способствует росту импакт-фактора получившего эти данные исследователя.

Рассмотрим более конкретный случай, связанный с предложением российского научного сообщества выполнить пилотную фазу проекта "Протеом человека" на примере идентификации белков 18-й хромосомы человека [14]. Казалось бы, такая формалистичная постановка задачи автоматически исключает из рассмотрения те коллективы, в исследования которых не входят белки данной хромосомы. Однако, как показано на рисунке 2, основанные на знаниях технологии позволяют определить отношение самых разных исследовательских проектов к генам 18-й хромосомы.

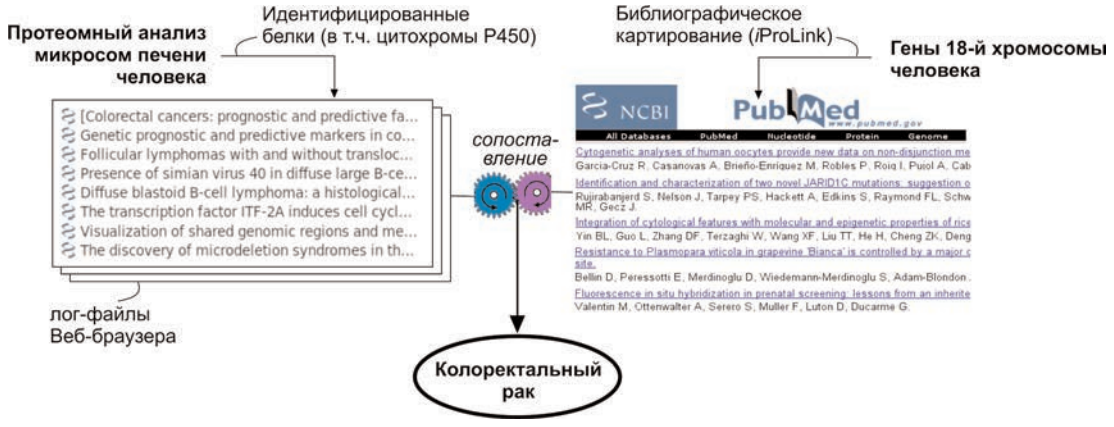

Рисунок 2.

Пример использования подхода, основанного на знаниях. Анализ лог-файлов сотрудников, проводивших обработку результатов идентификации белков микросомальной фракции печени, позволил выявить пять белков 18-хромосомы человека, ассоциированных с развитием колоректального рака. Сбор и статистическая обработка данных производилась с использованием автоматизированной системы BiblioEngine (ProContent, Германия).

На рисунке 2 рассмотрена область компетенции научной группы, работающей в области протеомики мембранной фракции тканей печени. Естественно, основным объектом исследования являются белки надсемейства цитохромов P450, катализирующие метаболизм лекарственных препаратов.

Заметим, что ни один белок из этого надсемейства не кодируется генами, расположенными в составе 18-й хромосомы. Был проведен анализ литературных источников, обращения к которым были зафиксированы в лог-файлах Веб-браузеров на служебных компьютерах сотрудников научной группы. Встретившиеся в лог-файлах идентификаторы рефератов, обращения к которым производились посредством сервера PubMed, сравнили с библиографическими картами, полученными для всех белков 18-й хромосомы с помощью сервера iProLink [15]. В результате было выявлено, что наибольшее количество совпадений между записями лог-файлов и библиографическими картами характерно для статей, связанных с колоректальным раком, поскольку на хромосоме №18 локализованы 5 генов, ассоциированных именно с этим заболеванием.

Объяснение вышеприведенного примера состоит в том, что биообразцы тканей печени наиболее часто поступают к исследователям после операций по поводу метастазирования колоректального рака в печень [16]. Анализируя результаты идентификации цитохромов P450, сотрудники научной группы постоянно обращали внимание на статьи, в которых их коллеги исследовали связь между метаболизмом химиотерапевтических средств и процессом метастазирования. Проведенный анализ позволяет включить в состав дальнейших экспериментов по протеомному профилированию печени белки 18-й хромосомы, связанные с колоректальным раком. Их идентификация уже будет производиться целенаправленным образом с использованием селективного мониторинга соответствующих пептидных ионов.

4. ИНИЦИАТИВНЫЙ ПРОЕКТ.

Препятствием для активного применения подхода, использующего базы знаний, в настоящее время является отсутствие надлежащим образом собранных данных об обращении пользователей к молекулярно-биологическим ресурсам. Формально информация зафиксирована на протоколирующих серверах Интернет-провайдеров, но использование этой информации затруднительно в силу её разрозненности. Эта информация находится в распоряжении различных организаций. Альтернативным вариантом получения такой информации является добровольное участие пользователей в ее накоплении.

Для накопления объема информации, нужного для пилотной демонстрации возможностей базы знаний организуется инициативный проект "Open Knowledge Initiative" [17]. В проекте может принять участие любой исследователь на условиях информированного согласия о предоставлении персональных данных для некоммерческого использования. Цель пилотного проекта - показать, что сведения об обращениях исследователей к молекулярно-биологическим ресурсам поддаются алгоритмической интерпретации. Ожидается, что результаты алгоритмической интерпретации должны совпасть с известными данными о строении и функциях молекулярных систем.

В настоящее время инициатива нашла поддержку в ведущих ВУЗах России, включая профильные факультеты и кафедры Московского государственного университета, Московской академии тонкой химической технологии, Сибирского государственного медицинского университета (г. Томск), Новосибирского госуниверситета и др. Участие в сотрудничестве заключается в инсталляции дополнительной панели (toolbar) в Интернет-браузере, с помощью которой осуществляется мониторинг обращений к фиксированному перечню молекулярно-биологических ресурсов. Более подробная информация об инициативе по свободному распространению знаний размещена на сайте www.okn-ex.ru.

Присоединиться к инициативе приглашаются исследователи, научные группы и организации, активно работающие с молекулярно-биологическими ресурсами. Важно учитывать, что чем больше экспертов участвуют в таком проекте, тем ценнее и качественнее данные, которые они получают из базы знаний. После набора определенной критической массы пользователей, польза, которую получает каждый отдельный пользователь, становится существенной. Накопленные данные будут предоставляться всем участникам проекта, как в

исходном виде, так и в статистически обработанной форме. По накоплению достаточного объема данных результаты будут опубликованы и предоставлены для оценки комитету по стандартизации международной организации "Протеом человека". Выполнение пилотного исследования позволит определить оптимальные пути дальнейшего развития баз знаний как эффективного инструмента организации научно-исследовательской и научно-образовательной деятельности в области технологий живых систем.

Авторы выражают благодарность специалистам, принявшим участие в обсуждении концепции постгеномных баз знаний, в том числе: председателю комитета HUPO по стандартизации в протеомике Хеннигу Хермъякобу (European Bioinformatics Institute, UK), со-председателю проекта по протеомике тканей мозга проф. Хельмуту Майеру (Medical Proteome Center, Germany), представителю ЕС по вопросам международного научно-технологического сотрудничества доктору Индриди Бенедиктссону, советнику Центрального банка РФ, проф. Щербакову А.Ю., а также коллегам, активно поддержавшим инициативу по развитию баз знаний в России: академику РАМН Арчакову А.И., академику РАМН Швецу В.И., академику РАН Ткачуку В.А., проректору СибГМУ проф. Рязанцевой Н.В., декану МБФ СибГМУ, проф. Карасю С.И., зам. зав. кафедрой биоинженерии МГУ проф. Шайтану К.В., зам. декана Медицинского факультета НГУ Пустыльняку В.О.

В работе использовались данные персональных лог-файлов сотрудников лаборатории микросомального окисления Института биомедицинской химии имени В.Н. Ореховича РАМН (рук. - д.б.н. Карузина И.И.). Программное обеспечение BiblioEngine любезно предоставлено компанией "КуБ" на условиях пробного тестирования (free trial).

При поддержке Российского фонда фундаментальных исследований (грант 09-04-12175-офи_м) и Рособразования (государственный контракт №П2215).

ЛИТЕРАТУРА

1. Entrez, The Life Science Search Engine, http://www.ncbi.nlm.nih.gov/Entrez/
2. UniProtKB, Protein Knowledgebase, http://www.uniprot.org/
3. Resource for Studying Biological Molecules, http://www.rcsb.org/
4. The GeneOntology, http://www.geneontology.org/
5. KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/
6. PRIDE: PRoteomics IDEntifications database, http://www.ebi.ac.uk/pride/
7. *Witbrock M., Matuszek C., Brusseau A., Kahlert R.C., Fraser C.B., Lenat D.* (2005) in: Papers from the 2005 AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (KCVC). Stanford, California, pp. 99-105.
8. *Кузнецов С.В.* (2004) Проблемы теории и практики управления, **6**, 85-89.
9. *Page L.* (2001) Method for node ranking in a linked database, US Patent 6285999.
10. *Vise D., Malseed M.* (2006) The Google Story, Дельта, М.
11. NCBI News, Featured Resource: An Expanded Set of Discovery Components in the Entrez System (2009) http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=newsncbi&part=JFKPFT637.bid.1
12. NLM Privacy Policy, http://www.nlm.nih.gov/privacy.html
13. The HUPO Proteomics Standards Initiative, http://www.psidev.info/
14. *Archakov A., Bergeron J.J., Khlunov A., Lisitsa A., Paik Y.K.* (2009) Mol Cell Proteomics, **8**(9), 2199-200.
15. *Hu Z.Z., Mani I., Hermoso V., Liu H., Wu C.H.* (2004) Comput. Biol. Chem., **28**(5-6), 409-416.
16. *Lisitsa A.V., Petushkova N.A., Thiele H., Moshkovskii S.A., Zgoda V.G, Karuzina I.I., Chernobrovkin A.L., Skipenko O.G, Archakov A.I.* (2009) J. Proteome Res., in press.

Поступила: 04. 11. 2009.

KNOWLEDGEBASES IN POSTGENOMIC MOLECULAR BIOLOGY

A.V. Lisitsa[1], B.V. Shilov[2], P.A. Evdokimov[1], S.A. Gusev[1,3]

[1]Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Pogodinskaya, 10, Moscow, 119121 Russia; tel.: +7-499-2463731; fax: +7-499-2450857; e-mail: fox@ibmh.msk.su
[2]Siberian State Medical University
[3]"Cryptome Research", Ltd

Knowledgebases can become an effective tool essentially raising quality of information retrieval in molecular biology, promoting the development of new methods of education and forecasting of the biomedical R&D. Knowledge-based technologies should induce "paradigm shift" in the life science due to integrative focusing of research groups towards the challenges of postgenomic era. This paper debates concept of the knowledgebase, which exploits web usage mining to personalize the access of molecular biologist to the Internet resources.

Key words: knowledgebase, formalization of knowledge, pertinent information, mass-spectrometry, human proteome.